Course Introduction

CS 4319

Data Mining & Warehouses

Dr. Oner Ulvi Celepcikay

Data Mining: Introduction

Chapter 1

Introduction to Data Mining

Data Mining: Introduction

- Session 1 → 5:30 – 6:40 pm
- Break → 6:40 – 7:00 pm
- Session 2 → 7:00 – 8:15 pm

Data Mining: Introduction

- Introduction
- Survey: www.PollEv.com/onercelepcik738
- Syllabus
- Writing Credits Students
- Teams (4-6 students)
- PDF of slides will be posted after class
- https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/
- http://www.forbes.com/sites/gregoryferenstein/2016/01/20/report-why-data-scientist-is-the-best-job-to-pursue-in-2016/#6253adc55f4b

Why Mine Data? Commercial Viewpoint

- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - purchases at department/ grocery stores (RFIDs)
  - Bank/Credit Card transactions
  - Customer Behavior, history

- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
  - Provide better, customized services for an *edge* (e.g. in Customer Relationship Management)

Why Would a Business Mine Data? One example

Kashmir Hill
Forbes Staff

FOLLOW

Welcome to The Not-So Private Parts where technology & privacy collide
full bio →

TECH 2/16/2012 @ 11:02AM | 2,649,316 views

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

+ Comment Now + Follow Comments

Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what you

http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/

One Recent Example

- So Target started sending coupons for baby items to customers according to their pregnancy scores.
- An angry man went into a Target outside of Minneapolis, demanding to talk to a manager:

"My daughter got this in the mail!" he said. "She's still in high school, and you're sending her coupons for baby clothes and cribs? Are you trying to encourage her to get pregnant?"

The manager looked at the mailer; it was addressed to the man's daughter and contained advertisements for maternity clothing, nursery furniture and pictures of smiling infants.

What did TARGET do to use this incident as an opportunity?

Why Mine Data? One recent example

Target assigns every customer a Guest ID number, tied to their credit card, name, or email address that becomes a bucket that stores a history of everything they've bought and any demographic information Target has collected from them or bought from other sources. Using that, Pole looked at historical buying data for all the ladies who had signed up for Target baby registries in the past. From the NYT:

> [Pole] ran test after test, analyzing the data, and before long some useful patterns emerged. Lotions, for example. Lots of people buy lotion, but one of Pole's colleagues noticed that women on the baby registry were buying larger quantities of unscented lotion around the beginning of their second trimester. Another analyst noted that sometime in the first 20 weeks, pregnant women loaded up on supplements like calcium, magnesium and zinc. Many shoppers purchase soap and cotton balls, but when someone suddenly starts buying lots of scent-free soap and extra-big bags of cotton balls, in addition to hand sanitizers and washcloths, it signals they could be getting close to their delivery date.

Why Mine Data? One recent example

> As Pole's computers crawled through the data, he was able to identify about 25 products that, when analyzed together, allowed him to assign each shopper a "pregnancy prediction" score. More important, he could also estimate her due date to within a small window, so Target could send coupons timed to very specific stages of her pregnancy.
>
> One Target employee I spoke to provided a hypothetical example. Take a fictional Target shopper named Jenny Ward, who is 23, lives in Atlanta and in March bought cocoa-butter lotion, a purse large enough to double as a diaper bag, zinc and magnesium supplements and a bright blue rug. There's, say, an 87 percent chance that she's pregnant and that her delivery date is sometime in late August.
>
> via How Companies Learn Your Secrets – NYTimes.com.

And perhaps that it's a boy based on the color of that rug?

Social Media Data Mining Example

Facebook Knows You Better Than Anyone Else

JAN. 19, 2015

Karen Bleier/Agence France-Presse — Getty Images

Think your friends know you well? Researchers have developed a computer model that can judge someone's personality more accurately than their friends and family — using nothing but the subject's Facebook activity.

Researchers at the University of Cambridge and Stanford University tested their algorithm on more than 17,000 Facebook users, who completed a personality survey and provided the researchers with access to their "likes." Many of their friends, colleagues and family members also completed a survey

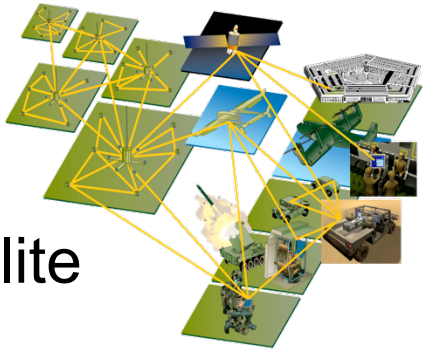
http://www.nytimes.com/2015/01/20/science/facebook-knows-you-better-than-anyone-else.html

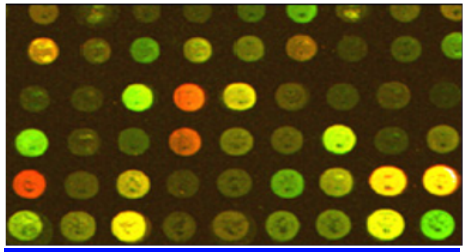
Social Media Data Mining Example

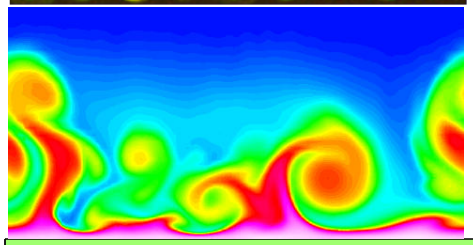
- The survey rated each subject on 5 OCEAN personality traits:
  - openness,
  - conscientiousness,
  - extraversion,
  - agreeableness and
  - neuroticism.
- Compared results to the subjects' Facebook activity to establish links between "likes" and specific personality traits.
- Given enough data***, the algorithm was better able to predict a person's personality traits than any of the human participants.*** It needed access to just 10 likes to beat a work colleague, 70 to beat a roommate, 150 to beat a parent or sibling, and 300 to beat a spouse.

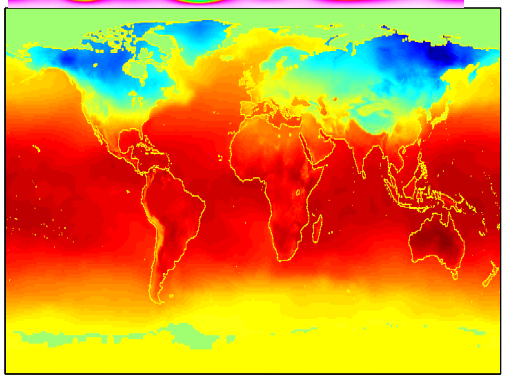
Why Mine Data? Scientific Viewpoint

- Data collected and stored at enormous speeds (GB/hour)
  - remote sensors on a satellite
  - telescopes scanning the skies
  - microarrays generating gene expression data
  - scientific simulations generating terabytes of data
- Traditional techniques infeasible for raw data
- Data mining may help scientists
  - in classifying and segmenting data
  - in Hypothesis Formation









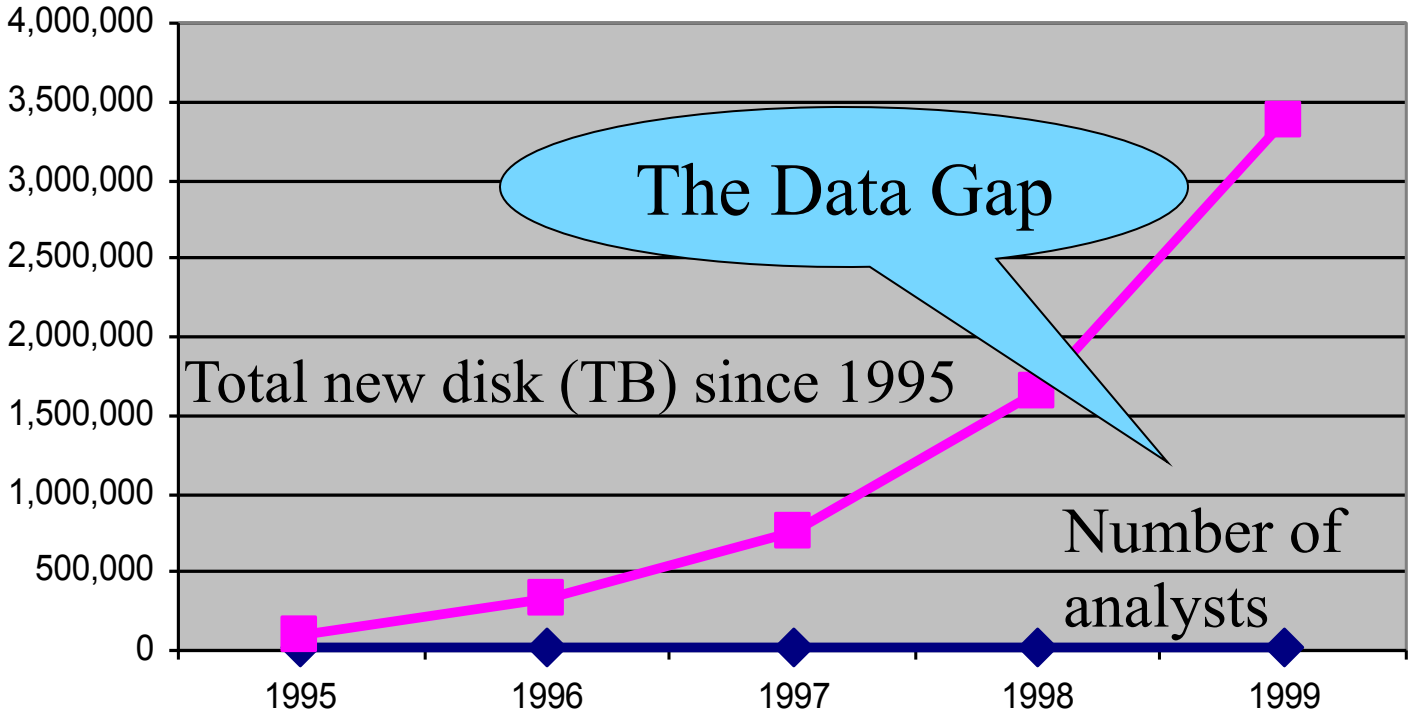
Data Mining: Introduction

- **Data Revolution:**

https://www.youtube.com/watch?v=0JaDiBiAnv8

Mining Large Data Sets - Motivation

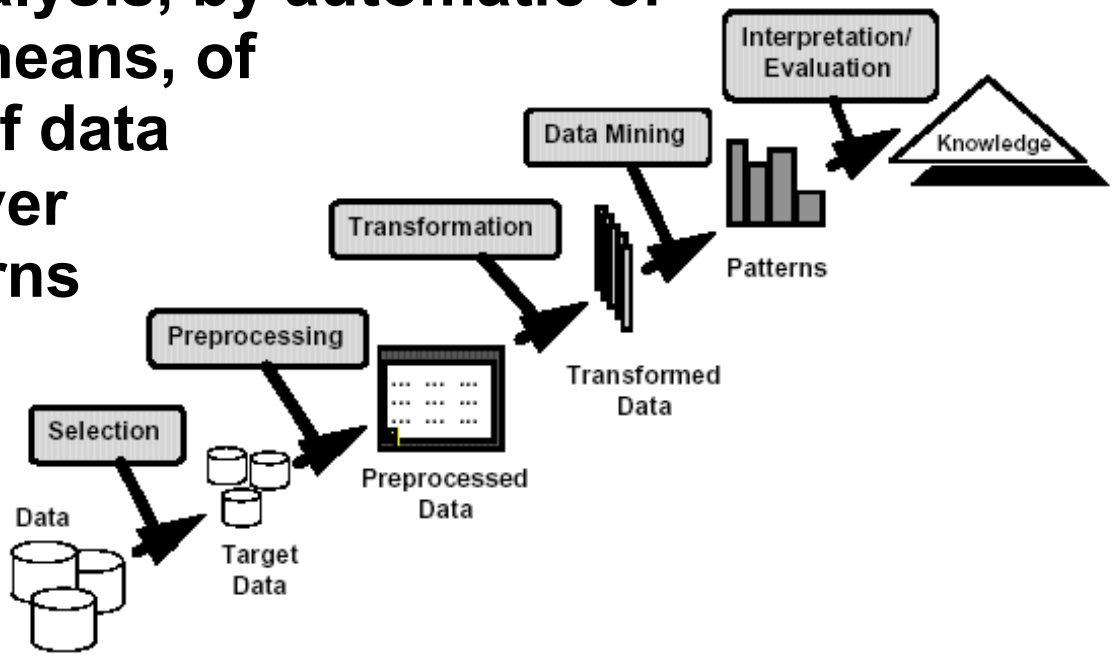
- There is often information "hidden" in the data that is not readily evident
- Human analysts may take weeks to discover useful information
- Much of the data is never analyzed at all



From: R. Grossman, C. Kamath, V. Kumar, "Data Mining for Scientific and Engineering Applications"

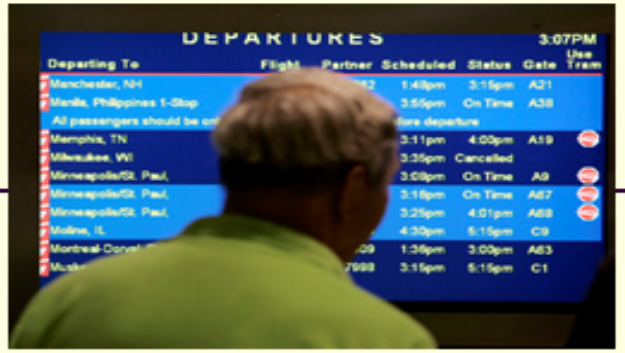
What is Data Mining?

- **Many Definitions**
  - **Non-trivial extraction of implicit, previously unknown and potentially useful information from data**
  - **Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns**



What is Next?

Data



Information

Knowledge

Wisdom

Next?

KDD & Data Mining

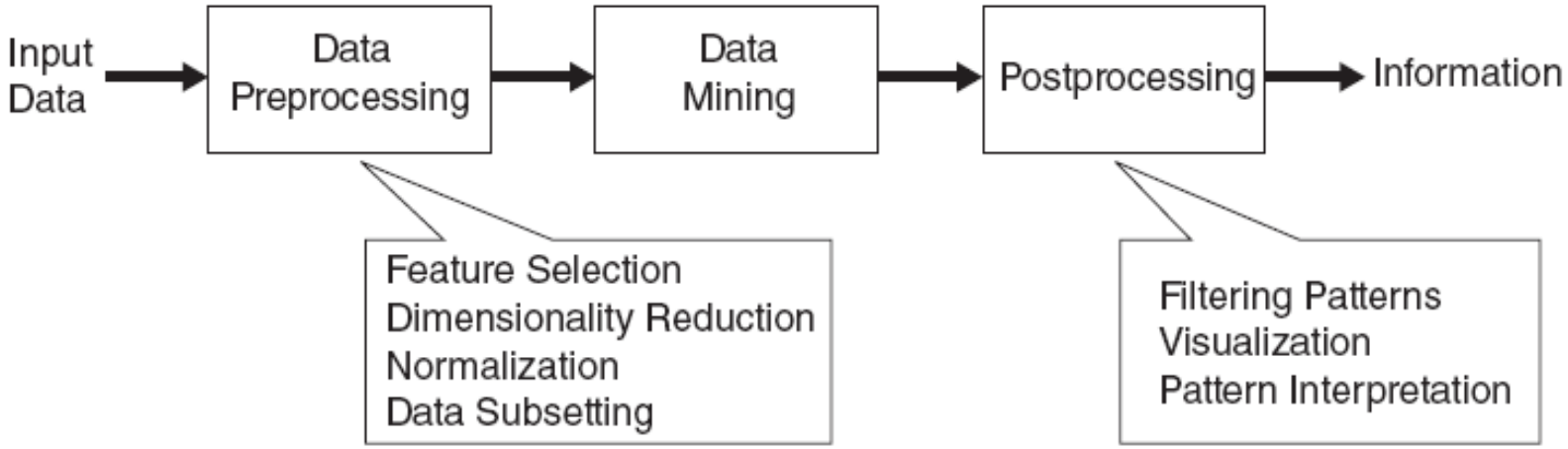


Figure 1.1. The process of knowledge discovery in databases (KDD).

What is (not) Data Mining?

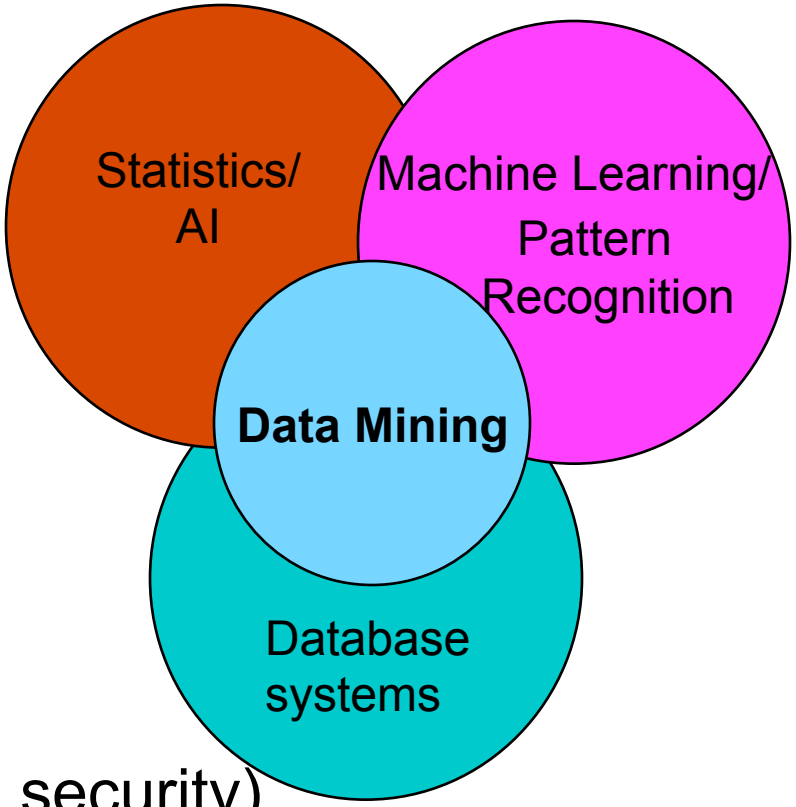
- **What is not Data Mining?**
  - Look up phone number in phone directory
  - Query a Web search engine for information about "Amazon"

- **What is Data Mining?**
  - Certain names are more prevalent in certain US locations (O'Brien, O'Rurke, O'Reilly… in Boston area)
  - Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com,)

Origins of Data Mining

- Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems
- Traditional Techniques may be unsuitable due to
  - Enormity of data
  - High dimensionality of data
  - Heterogeneous, distributed nature of data

  (Reducing Com., consolidation, security)



Data Mining Tasks

- Prediction Methods
  - Use some variables to predict unknown or future values of other variables.

  e.g. house prices, medical studies, dep./ind. variables

- Description Methods
  - Find human-interpretable patterns that describe the data.

  e.g.: correlation, clusters, trajectories, anomalies)

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

Data Mining Tasks...

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation Detection [Predictive]

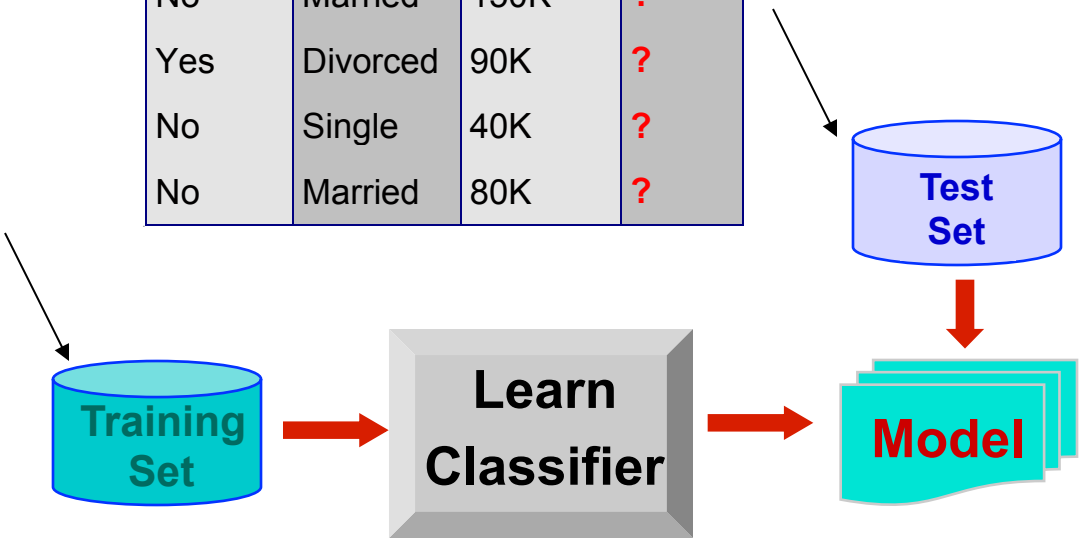
Classification: Definition

- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

Classification Example

Tid	Refund (categorical)	Marital Status (categorical)	Taxable Income (continuous)	Cheat (class)
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



Classification: Application 1

- Direct Marketing
  - Goal: Reduce cost of mailing by *targeting* a set of consumers likely to buy a new cell-phone product.
  - Approach:
    - Use the data for a similar product introduced before.
    - We know which customers decided to buy and which decided otherwise. This *{buy, don't buy}* decision forms the *class attribute*.
    - Collect various demographic, lifestyle, and company-interaction related information about all such customers.
      - Type of business, where they stay, how much they earn, etc.
    - Use this information as input attributes to learn a classifier model.

From [Berry & Linoff] Data Mining Techniques, 1997

Classification: Application 2

- Fraud Detection
  - Goal: Predict fraudulent cases in credit card transactions.
  - Approach:
    - Use credit card transactions and the information on its account-holder as attributes.
      - When does a customer buy, what does he buy, how often he pays on time, etc
    - Label past transactions as fraud or fair transactions. This forms the class attribute.
    - Learn a model for the class of the transactions.
    - Use this model to detect fraud by observing credit card transactions on an account.

Classification: Application 3

- Customer Attrition/Churn:
  - Goal: To predict whether a customer is likely to be lost to a competitor.
  - Approach:
    - Use detailed record of transactions with each of the past and present customers, to find attributes.
      - How often the customer calls, where he calls, what time-of-the day he calls most, his financial status, marital status, etc.
    - Label the customers as loyal or disloyal.
    - Find a model for loyalty.

From [Berry & Linoff] Data Mining Techniques, 1997

Classification: Application 4

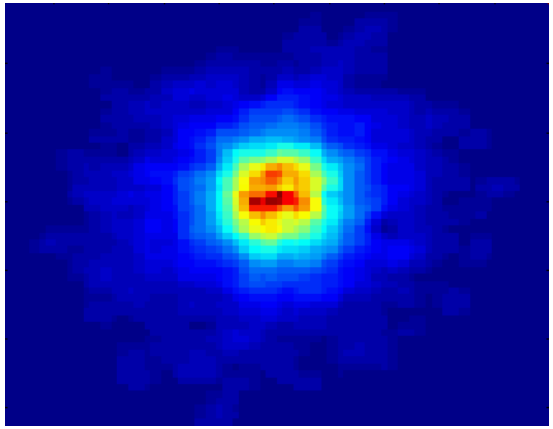
- Sky Survey Cataloging
  - Goal: To predict class (star or galaxy) of sky objects, especially visually faint ones, based on the telescopic survey images (from Palomar Observatory).
    - 3000 images with 23,040 x 23,040 pixels per image.
  - Approach:
    - Segment the image.
    - Measure image attributes (features) - 40 of them per object.
    - Model the class based on these features.
    - Success Story: Could find 16 new high red-shift quasars, some of the farthest objects that are difficult to find!

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

Classifying Galaxies

Courtesy: http://aps.umn.edu

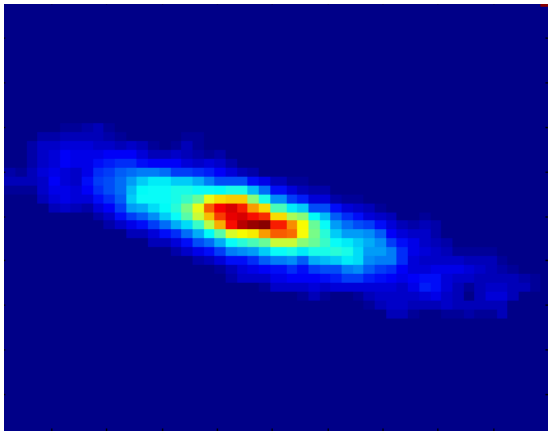
Early



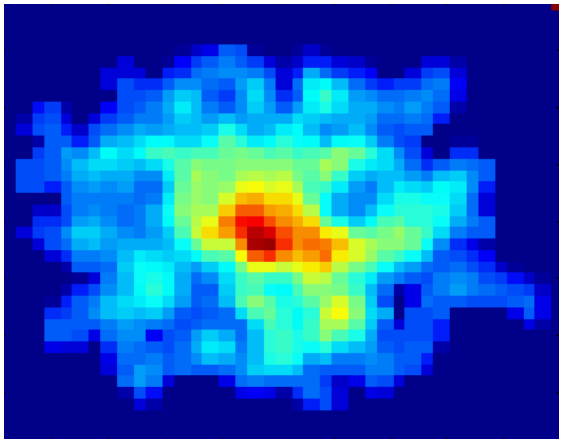
Class:
- **Stages of Formation**

Attributes:
- **Image features,**
- **Characteristics of light waves received, etc.**

Intermediate



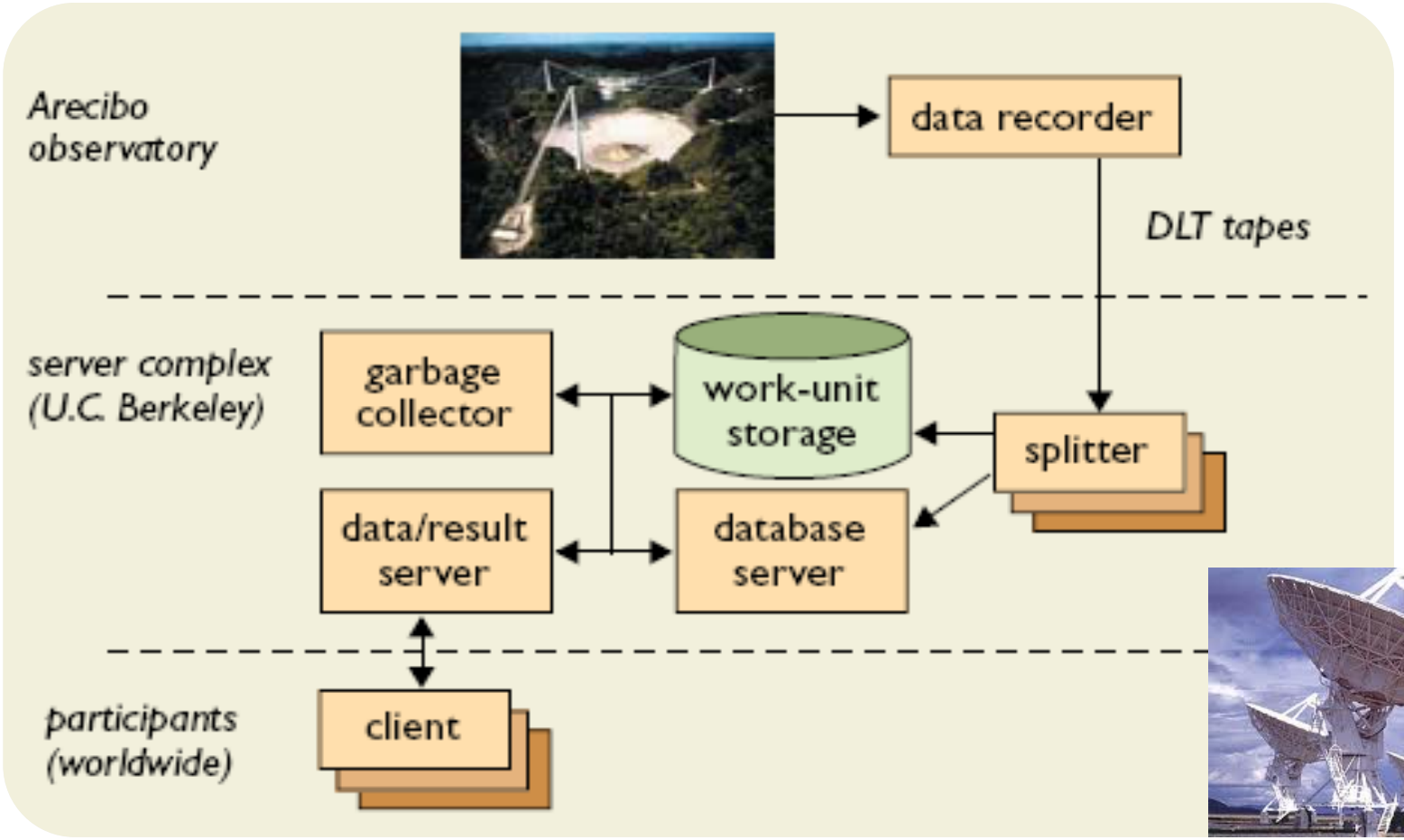
Late



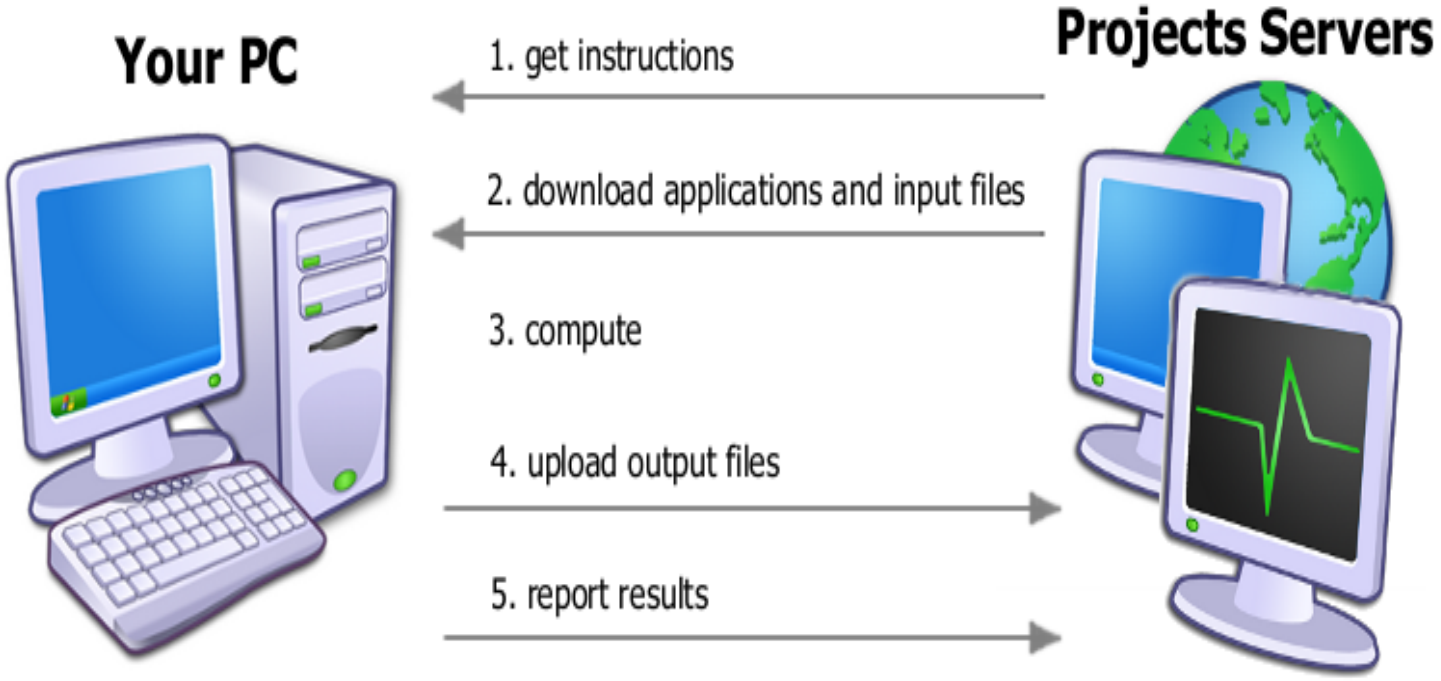
Data Size:
- **72 million stars, 20 million galaxies**
- **Object Catalog: 9 GB**
- **Image Database: 150 GB**

Another Example: SETI@Home Project

- 3.6 M users, 226 countries, 40TB data, 1M years CPU



Another Example: SETI@Home Project



An example of how to utilize Data

Human Computation

https://www.youtube.com/watch?v=Aszl5avDtek

Clustering Definition

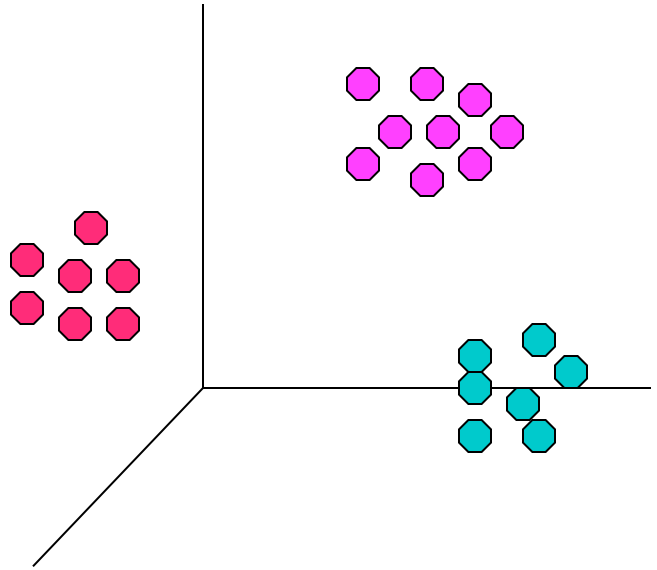
- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.
- Similarity Measures:
  - Euclidean Distance if attributes are continuous.
  - Other Problem-specific Measures.

Illustrating Clustering

- Euclidean Distance Based Clustering in 3-D space.

Intracluster distances are minimized

Intercluster distances are maximized



Clustering: Application 1

- Market Segmentation:
  - Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
  - Approach:
    - Collect different attributes of customers based on their geographical and lifestyle related information.
    - Find clusters of similar customers.
    - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

Clustering: Application 2

- Document Clustering:
  - Goal: To find groups of documents that are similar to each other based on the important terms appearing in them.
  - Approach: To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.
  - Gain: Information Retrieval can utilize the clusters to relate a new document or search term to clustered documents.

Illustrating Document Clustering

- Clustering Points: 3204 Articles of Los Angeles Times.
- Similarity Measure: How many words are common in these documents (after some word filtering).

Category	Total Articles	Correctly Placed
Financial	555	364
Foreign	341	260
National	273	36
Metro	943	746
Sports	738	573
Entertainment	354	278

Association Rule Discovery: Definition

- Given a set of records each of which contain some number of items from a given collection;
  - Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

Association Rule Discovery: Application 1

- Marketing and Sales Promotion:
  - Let the rule discovered be

    {Bagels, … } --> {Potato Chips}
  - Potato Chips as consequent => Can be used to determine what should be done to boost its sales.
  - Bagels in the antecedent => Can be used to see which products would be affected if the store discontinues selling bagels.
  - Bagels in antecedent *and* Potato chips in consequent => Can be used to see what products should be sold with Bagels to promote sale of Potato chips!

Association Rule Discovery: Application 2

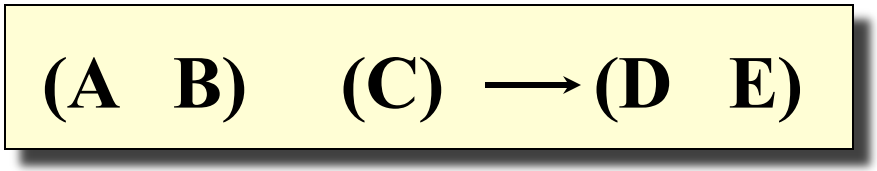
- Supermarket shelf management.
  - Goal: To identify items that are bought together by sufficiently many customers.
  - Approach: Process the point-of-sale data collected with barcode scanners to find dependencies among items.
  - A classic rule –(for young men)
    - If a customer buys diaper and milk, then he is very likely to buy beer.
    - So, don't be surprised if you find six-packs stacked next to diapers!

Association Rule Discovery: Application 3

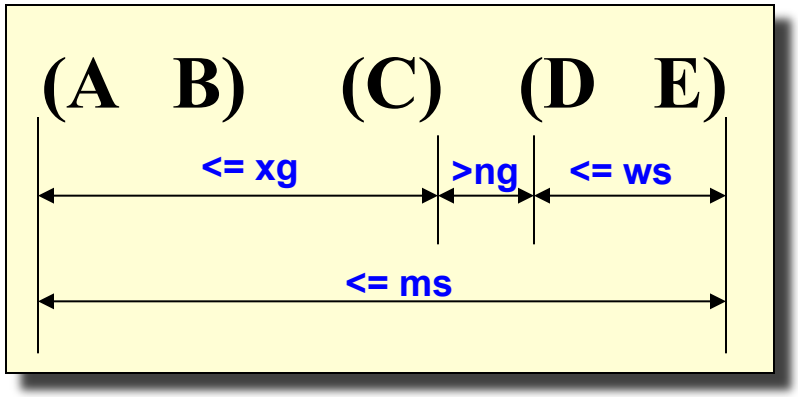
- Inventory Management:
  - Goal: A consumer appliance repair company wants to anticipate the nature of repairs on its consumer products and keep the service vehicles equipped with right parts to reduce on number of visits to consumer households.
  - Approach: Process the data on tools and parts required in previous repairs at different consumer locations and discover the co-occurrence patterns.

Sequential Pattern Discovery: Definition

- Given is a set of *objects*, with each object associated with its own *timeline of events*, find rules that predict strong sequential dependencies among different events.



- Rules are formed by first disovering patterns. Event occurrences in the patterns are governed by timing constraints.



Sequential Pattern Discovery: Examples

- In telecommunications alarm logs,
  - (Inverter_Problem Excessive_Line_Current)
    (Rectifier_Alarm) --> (Fire_Alarm)
- In point-of-sale transaction sequences,
  - Computer Bookstore:
    (Intro_To_Visual_C) (C++_Primer) -->
    (Perl_for_dummies,Tcl_Tk)
  - Athletic Apparel Store:
    (Shoes) (Racket, Racketball) --> (Sports_Jacket)

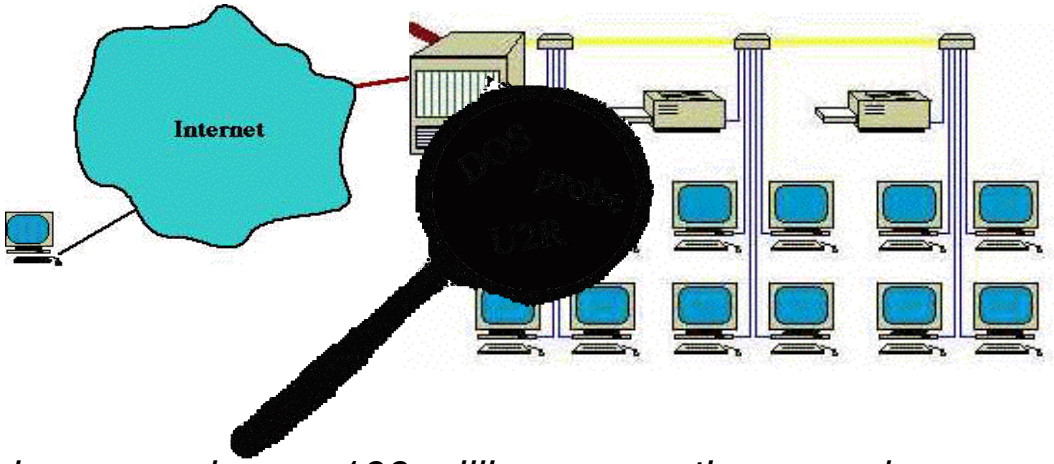
Regression

- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Greatly studied in statistics, neural network fields.
- Examples:
  - Predicting sales amounts of new product based on advertising expenditure.
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Time series prediction of stock market indices.

Deviation/Anomaly Detection

- Detect significant deviations from normal behavior
- Applications:
  - Credit Card Fraud Detection

  - Network Intrusion Detection



Typical network traffic at University level may reach over 100 million connections per day

Challenges of Data Mining

- Scalability
- Dimensionality
- Complex and Heterogeneous Data
- Data Quality
- Data Ownership and Distribution
- Privacy Preservation
- Streaming Data

Challenges of Data Mining

- Scalability
- Dimensionality
- Complex and Heterogeneous Data
- Data Quality
- Data Ownership and Distribution
- Privacy Preservation
- Streaming Data